CERTIFICATE OF MAILING BY "EXPRESS MAIL"

EXPRESS MAIL LABEL NUMBER: EL 617 042 686 US

DATE OF DEPOSIT: August 27, 2001

I HEREBY CERTIFY THAT THIS PAPER OR FEE IS BEING DEPOSITED WITH THE UNITED STATES POSTAL SERVICE "EXPRESS MAIL POST OFFICE TO ADDRESSEE" SERVICE UNDER 37 CFR 1.10 ON THE DATE INDICATED ABOVE AND IS ADDRESSED TO THE COMMISSIONER OF PATENTS, BOX PATENT APPLICATION, WASHINGTON, D.C. 20231.

MIKHAIL BAYLEY

UTILITY

APPLICATION

For

UNITED STATES LETTERS PATENT

on

METHOD FOR THE EVOLUTIONARY DESIGN OF BIOCHEMICAL REACTION NETWORKS

by

Bernhard O. Palsson

Jeremy S. Edwards

Sheets of Drawings: Seventeen (17)
Docket No.: UCSD1320-1

Attorneys

Gray Cary Ware & Freidenrich LLP
4365 Executive Drive, Suite 1600
San Diego, California 92121-2189

METHOD FOR THE EVOLUTIONARY DESIGN OF BIOCHEMICAL REACTION NETWORKS

[0001] This application claims priority under 35 USC 119(e) to United States provisional application Serial No. 60/265,554, filed January 31, 2001, the entire contents of which is incorporated herein by reference.

[0002] This invention was made in part with government support under Grant No. GM 57089 awarded by the National Institutes of Health. The government has certain rights in this invention.

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

[0003] The invention relates generally to biochemical reaction networks and more specifically to reconstruction of metabolic networks in an organism to obtain optimal desired whole cell properties.

BACKGROUND INFORMATION

[0004] Genome sequencing and annotation technologies are giving us detailed lists of the molecular components that cells are comprised of, and high-throughput technologies are yielding information about how these components are used. Thus we are approaching the stage where biological design is possible on a genome scale. It has proven difficult to 'splice' one gene from one organism into another and produce predictable results. The primary reason is that every component in a living cell has been honed through a lengthy evolutionary process to 'fit' optimally into the overall function of the cell. Simply introducing a foreign gene, or deleting an existing gene does not lead to predictable nor optimal results. Methods are needed to *a priori* predict the consequences of single or multiple gene deletions or additions on the function of an entire cellular function and force the remaining components to function in a predetermined manner. Such methods are lacking, although limited progress has been made with metabolic function on a cellular scale.

[0005] The interest in the redirection of metabolic fluxes for medical and industrial purposes has existed for some time. As a result of this interest, the field of metabolic engineering has been

born, and the primary goal of metabolic engineering is to implement desirable metabolic behavior in living cells. Advances and applications of several scientific disciplines including computer technology, genetics, and systems science lie at the heart of metabolic engineering.

[0006] The traditional engineering approach to analysis and design utilizes a mathematical or computer model. For metabolism this would require a computer model that is based on fundamental physicochemical laws and principles. The metabolic engineer then hopes that such models can be used to systematically 'design' a new and improved living cell. The methods of recombinant DNA technology should then be applied to achieve the desired cellular designs.

[0007] The 25-30 year history of metabolic analysis has demonstrated the need to quantitate systemic aspects of cellular metabolism, (*see e.g.*, Fell D., Understanding the control of metabolism, (London, Portland Press) (1996); Heinrich R., et al., Metabolic regulation and mathematical models, *Progress in Biophysics and Molecular Biology*, 32:1-82, (1977); Heinrich R. and Schuster S., The regulation of cellular systems, (New York, Chapman & Hall), xix, p. 372 (1996); Savageau M.A., Biochemical systems analysis. I. Some mathematical properties of the rate law for the ecomponent enzymatic reactions, *J. Theor. Biol.* 25(3):365-69 (1969)). There are significant incentives to study metabolic dynamics. A quantitative description of metabolism and the ability to produce metabolic change is not only important to achieve specific therapeutic goals but has general importance to our understanding of cell biology. Important applications include strain design for the production of therapeutics and other biochemicals, assessment of the metabolic consequences of genetic defects, the synthesis of systematic methods to combat infectious disease, and so forth. Quantitative and systemic analysis of metabolism is thus of fundamental importance. However, a review in the field has concluded that "despite the recent surge of interest in metabolic engineering, a great disparity still exists between the power of available molecular biological techniques and the ability to rationally analyze biochemical networks" (Stephanopoulos G., Metabolic engineering. *Current Opinions in Biotechnology*, 5:196-200 (1994)). Although this statement is a few years old, it still basically holds true. This conclusion is not surprising for we are competing with millions of years of natural evolution that achieves the best fitness of an organism in a given environment.

[0008] Although partial gene regulatory networks containing a small number of reactions have been designed (reviewed in Hasty et al., Computational studies of gene regulatory networks: *In numero* molecular biology, *Nature*, 2: 268-79 (2001)), the *a priori* design of biochemical regulatory networks, such as metabolic networks with defined performance characteristics and their subsequent construction has not been reduced to practice. The primary reason is that reliable detailed kinetic models cannot be constructed for an entire metabolic network, mainly because there are too many kinetic parameters whose numerical values must be determined and the detailed kinetic equations are by-and-large unknown. Thus, *a priori* design of optimal biochemical reaction networks, such as metabolism, is not possible because predictive kinetic models cannot be achieved. In fact, the values of the kinetic constants change with time due to mutations and an evolutionary process.

[0009] Heretofore it has been impossible to predict the end point of evolutionary processes as they are expected to be the outcome of the selection from random events. This invention discloses a method that allows for the *a priori* calculation of the endpoint of the evolution of metabolic networks in a defined environment. Although there are other mathematical modeling methods that are based on optimization principles in biological systems; i.e. the cybernetic modeling approach (Varner J. and Ramkrishna D., "Mathematical models of metabolic pathways," *Curr. Opin. Biotechnol.*, 10(2):146-50, (1999)), they are not amenable to the design of biological networks due to the number of parameters required. It thus gives the basis for the use of an evolutionary process to create or build such designs.

SUMMARY OF THE INVENTION

[0010] The present invention relates to a method for achieving an optimal function of a biochemical reaction network in a living cell. The biochemical reaction network can be a comprehensive biochemical reaction network, a substantially whole biochemical reaction network, or a whole biochemical reaction network. The method can be performed *in silico* using a reconstruction of a biochemical reaction network of a cell. The method can further include laboratory culturing steps to confirm and possibly expand the determinations made using the *in silico* steps, and to produce a cultured cell, or population of cells, with optimal functions.

[0011] The method can be performed by representing a listing of biochemical reactions in a network in a computer, such as by providing a database of biochemical reactions in a network; using optimization methods to calculate the optimal properties of the network; altering the list of reactions in the network and re-computing the optimal properties; and repeating the process described above until the desired performance is reached. The method may further include constructing the genetic makeup of a cell to contain the biochemical reactions which result from the optimization procedure; placing the cells constructed thereunder in culture under the specified environment; and cultivating the cells for a sufficient period of time under conditions to allow the cells to evolve to the determined desired performance.

[0012] The biochemical reaction network can be a metabolic network, for example a regulatory network. In addition, the cell whose genetic makeup is constructed can be a prokaryotic cell or a eukaryotic cell; such as *E. coli*, *S. cerevisiae,* chinese hamster ovary cells, and the like. Furthermore, the genetic makeup of a cell can be constructed by altering one or more genes in the cell, for example by addition or deletion, or by altering the regulation of a gene through its regulatory components (e.g., promoter, transcription factor binding sites, etc.). In another aspect, the invention provides an enriched population of cells produced by the method described above.

[0013] In another aspect, the present invention provides a method for achieving optimal functions of a comprehensive biochemical reaction network in a cell by providing a database including biochemical reactions in the network; using optimization methods to calculate the optimal properties of the network; receiving a user's selection for altering the reactions in the network and recomputing the optimal properties; repeating optimization until the desired property criterion is met; displaying the results of the optimization for constructing the genetic makeup of a cell so that it contains the biochemical reactions as a result of the optimization information; culturing the cells constructed under the specified environment conditions; and cultivating the cells for a sufficient period of time so that the cells evolve to the desired performance.

[0014] The optimization method may be carried out using a computer system provided by the present invention. The computer system typically includes a database that provides information

regarding one or more biochemical reaction networks of at least one organism; a user interface capable of receiving a selection of one or more biochemical reaction networks for optimization and/or comparison, and capable of receiving a selection of a desired performance; and a function for carrying out the optimization method calculations and recalculations. The computer system of the present invention can include a function for performing biochemical reaction network reconstruction. The database can be an internal database or an external database.

[0015] In another aspect the present invention provides a computer program product that includes a computer-usable medium having computer-readable program code embodied thereon. The program code is capable of interacting with the database and effects the following steps within the computing system: providing an interface for receiving a selection of a desired performance of the networks; determining the desired optimal properties, displaying the results of the determination, and altering the biochemical reaction network, before recalculating optimal properties of the biochemical reaction network, and repeating the process until a desired optimal function is achieved. Altering the biochemical reaction network can be performed based on an alteration manually input by a user, or can be performed automatically by the program code. The computer program can further provide an identification of database entries that are part of a reconstructed biochemical network, or can perform biochemical reaction network reconstruction.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Figure 1: The acetate uptake rate (AUR in units of mmole/g-DW/hr, g-DW is gram dry weight) versus oxygen uptake rate (OUR, in units of mmole/g-DW/hr) phenotype phase plane. The in silico defined line of optimality (LO) is indicated in the figure. The slope of this line is also indicated in the figure. The experimental data points are displayed on the figure. The error bars represent a single standard deviation, and the error bars are displayed for both the acetate and the oxygen uptake rate measurements. A linear regression was performed on the data points to define the experimentally reconstructed line of optimality. The correlation coefficient R^2 value for the curve fit is 0.92. Regions 1 & 2 represent distinct non-optimal metabolic phenotypes

[0017] Figure 2: The three-dimensional rendering of the phase surface for growth of E. coli on acetate. The x and y axis represent the same variables as in figure 1. The third dimension (the z-dimension) represents the cellular growth rate. The z-axis values are in gray scale with the optimal growth rate value quantitatively indicated on the corresponding legend. The line of optimality (LO) in three-dimensions is indicated. The parametric equation of LO in three-dimensions is indicated in the text. The black lines define the surface of the metabolic capabilities in the three-dimensional projection of the flux cone and represent constant values of the acetate uptake rate or oxygen uptake rate. The quantitative effect on cellular growth potential of increasing the acetate uptake rate (without proportional increase in the oxygen uptake rate) can be visualized. The data points are also plotted on the three-dimensional figure and error bars have been omitted.

[0018] Figure 3: Line of optimality for growth on acetate projected onto a plane formed by the acetate uptake rate and the growth rate. The data points have also been projected and a linear regression was performed in the two-dimensional plane to experimentally define the line of optimality. The line of optimality is indicated as a gray line and the regression line as a black line.

[0019] Figure 4: Line of optimality for growth on acetate projected onto a plane formed by the oxygen uptake rate and the growth rate. The data points have also been projected and a linear regression was performed in the two-dimensional plane to experimentally define the line of optimality. The line of optimality is indicated as a gray line and the regression line as a black line.

[0020] Figure 5: The succinate uptake rate (mmole/g-DW/hr) versus oxygen uptake rate (mmole/g-DW/hr) represented in the phenotype phase plane. The labeled line is the *in silico* defined line of optimality (LO). The experimental data points are displayed on the figure. The error bars are displayed for both the succinate and oxygen uptake rate measurements and represent a single standard deviation. Cultivations for which acetate was produced above a threshold of 0.3 mmole / gDW / hr are indicated by open circles, filled circles identify either no acetate production or production below the threshold. The black dotted line represents the linear regression of the data points with no acetate production.

[0021] Figure 6: The measured acetate production vs. the in silico predictions for each point illustrated in figure 5. The data points are rank ordered by the magnitude of the succinate uptake rate.

[0022] Figure 7: Three-dimensional phenotype phase plane for E. coli growth on succinate. The x and y axis represent the same variables as in figure 6. The third dimension (the z-axis) represents the cellular growth rate. The z-axis values are in gray scale with the corresponding legend in the figure. The demarcation lines separating the colored regions represent constant oxygen and acetate uptake rates, and the quantitative effect of moving away from the line of optimality can be visualized. The data points are plotted in this three-dimensional figure with the exception of the error bars.

[0023] Figure 8: Calculated and experimental values for growth of *E. coli* K-12 on glucose. The glucose uptake rate, GUR (mmole/gDW/h), and the oxygen uptake rate, OUR (mmole/gDW/h), are shown in the phenotype phase plane. The LO is indicated. Data points are confined to the LO or the acetate overflow region, where acetate secretion is predicted *in silico* and experimentally observed.

[0024] Figure 9: Three-dimensional rendering of growth rates graphed over the phase plane for growth on glucose. The x and y axes represent the same variables as in fig 8. The z-axis represents the cellular growth rate, with color-coded values and the optimal growth rate indicated on the legend.

[0025] Figure 10: GUR plotted against OUR with experimental values for adaptive evolution experiments. Data points lie near the LO and in region 2 where acetate overflow is predicted.

[0026] Figure 11: Three-dimensional rendering of the post-evolutionary 3D growth surface over the glucose phenotypic phase plane. All data points cluster tightly on or near the LO.

[0027] Figure 12: Calculated and experimental values for growth on glycerol. the glycerol uptake rate, GlUR (mmole/gDW/h), and the oxygen uptake rate, OUR (mmole/gDW/h), are shown in the phenotype phase plane. The LO is shown in indicated. The experimental data points are confined to region 1, characterized by futile cycles and suboptimal growth rates.

[0028] Figure 13: Three-dimensional rendering of growth rates graphed over the glycerol phenotypic phase plane. The x and y axes represent the same variables as in figure 12. The z-axis represents the cellular growth rate, and the optimal growth rate indicated on the legend. No data points lie near the LO.

[0029] Figure 14: The glycerol uptake rate (GlUR) plotted against the oxygen uptake rate (OUR) with experimental values for adaptive evolution experiments. The starting point of evolution is indicated (day 0). Experimental values for the first evolutionary trajectory (E1) are indicated in blue, while values for the second evolutionary trajectory (E2) are indicated in green. In both experiments, the initial strain converges towards a similar endpoint on the LO, representing optimal growth rates.

[0030] Figure 15: Change in growth rate in units or hr^{-1}, with time for adaptive evolution experiments on glycerol. Both experiments reveal a similar adaptation profile, with increased fitness and a doubling of the growth rate.

[0031] Figure 16: The glycerol-oxygen phenotypic phase plane with experimental values for growth on glycerol after the adaptive evolution experiments. All values cluster tightly on or near the LO. Data represents the titration of the carbon source and the quantitative effect of moving along the LO.

[0032] Figure 17 Three-dimensional rendering of the post-evolutionary phase surface. All data points cluster tightly on or near the LO.

DETAILED DESCRIPTION OF THE INVENTION

[0033] One aspect of this invention provides a method to design the properties of a large biochemical reaction network. Using a method of this aspect of the present invention, a biochemical reaction network can be designed to a predetermined performance in a specified environment. One aspect of the invention includes:

[0034] 1) using a computer reconstruction of the reaction structure of a biochemical reaction network,

[0035] 2) using optimization methods to determine the optimal functionalities of the reaction network,

[0036] 3) changing the reaction structure in the computer representation of the network by removing or adding a single or a multitude of genes and recalculating the optimal properties,

[0037] 4) using genetic manipulations to get the gene complement in an organism to correspond to the structure of the reaction network whose optimal properties have been determined through computer simulation, and

[0038] 5) using extended cultivation under a defined selection pressure to evolve the function of the actual reaction network toward the optimal solution that was predicted by the computer simulation. The adaptive evolutionary process will itself determine the best set of kinetic parameters to achieve the optimal design. More than one similar set of parameter values can be determined thorough the evolutionary process.

[0039] Using the methods and procedures disclosed herein, a biochemical reaction network can be designed *a priori* in a computer. Following the design of the reaction network, an evolutionary process is carried out in the laboratory under the appropriate conditions on a genetically modified organism or a wild-type strain that corresponds to the network used for the computer simulations. Organisms may achieve the optimal behavior in a non-unique fashion--that is there may be equivalent optimal solutions. Thus the invention involves a non-obvious and non-existing combination of computer design methods, genetic alteration, and evolutionary process to achieve optimal performance of biochemical reaction networks within the environment of a living cell.

[0040] In another aspect, the present invention relates to a method for determining optimal functions of a comprehensive biochemical reaction network in a living cell. The method is used to achieve a desired performance of the living cell. The method can be performed by representing a listing of the biochemical reactions in the network in a computer; using optimization methods to calculate the optimal properties of the network; altering the list of reactions in the network and re-computing the optimal properties; and repeating the altering step until the desired performance is met.

[0041] In addition to the above steps which are performed *in silico,* the method can further include steps involving culturing a living cell, or a population of cells. These steps include constructing the genetic makeup of a cell to contain the biochemical reactions which result from repeating the altering step until the desired performance are met; placing the cell constructed thereunder in culture under the specified environment; and cultivating the cell for a sufficient period of time and under conditions to allow the cell to evolve to the determined desired performance.

[0042] A biochemical reaction network is an interrelated series of biochemical reactions that are part of a biochemical pathway or linked biochemical pathways. Many biochemical reaction networks have been identified such as metabolic reaction networks, catabolic reaction networks, polypeptide and nucleic acid synthesis reaction networks, amino acid synthesis networks, energy metabolism and so forth. Other types of biochemical reaction networks include regulatory networks including cell signaling networks, cell cycle networks, genetic networks involved in regulation of gene expression, such as operon regulatory networks, and actin polymerization networks that generate portions of the cytoskeleton. Most of the major cell functions rely on a network of interactive biochemical reactions.

[0043] To practice the present invention, the reaction structure of a comprehensive, preferably substantially whole, or most preferably whole biochemical reaction network in an organism to be biochemically designed must be reconstructed for computer simulations. A whole biochemical reaction network includes all of the biochemical reactions of a cell related to a certain biochemical function. For example a whole metabolic reaction network includes essentially all of the biochemical reactions that provide the metabolism of a cell. Metabolic reaction networks exemplify a universal biochemical reaction network found in some form in all living cells.

[0044] A comprehensive biochemical reaction network is an interrelated group of biochemical reactions that effect a detectable property, and that can be modified in a predictable manner with respect to the effect of such modifications on the detectable property in the context of a living cell. For example, a comprehensive biochemical reaction network can include core reactions that effect the yield of a biomolecule produced by the cell, even though the core reactions include only a portion of the reactions in the whole biochemical reaction network involved in yield of the

biomolecule, provided that computational methods can be used to predict the effect of changes in the core biochemical reactions on the yield in a living cell.

[0045] A substantially whole biochemical reaction network is an interrelated group of biochemical reactions that are responsible for a detectable property of a living cell. Substantially whole biochemical reaction networks include core reactions as well as secondary reactions that have an effect on the detectable property, even though this effect can be relatively minor. Changes in substantially whole biochemical reaction networks can be predicted using computational methods. The present invention can also utilize the majority of reactions in a whole biochemical reaction network, rather than a comprehensive, substantially whole, or whole biochemical reaction network.

[0046] Optimal properties, also referred to herein as optimal functions, determined using the methods of the current invention include, for example, glycerol uptake rate, oxygen uptake rate, growth rate, sporulation occurrence and/or rates, rates of scouring of rare elements under nutritionally poor conditions, biomass, and yields of biomolecules such as proteins, carbohydrates, antibiotics, vitamins, amino acids, fermentation products, such as lactate production. Optimal properties also include, for example, yields of chiral compounds and other low molecular weight compounds. Optimal properties also include, for example, the maximal internal yields of key co-factors, such as energy carrying ATP or redox carrying NADPH and NADH. Optimal properties can also be defined by a cellular engineer to include properties such as flux rates through key reactions in the biochemical reaction network. The current invention allows an optimal performance related to one or more of these properties to be achieved. For example, the methods allow a specific desired growth rate or specific desired yield to be achieved.

[0047] Typically for the methods of the current invention, the biochemical reactions of a reconstructed biochemical reaction network are represented in a computer. This representation can include a listing of the biochemical reactions in the reconstructed biochemical reaction network. The listing can be represented in a computer database, for example as a series of tables of a relational database, so that it can be interfaced with computer algorithms that represent network simulation and calculation of optimality properties.

[0048] The biochemical network reconstruction must be of high quality for the present invention. The process of high quality biochemical reaction network, specifically metabolic reaction network, reconstruction has been established M.W. Covert, C.H. Schilling, I. Famili, J.S. Edwards, I.I. Goryanin, E. Selkov, and B.O. Palsson, "Metabolic modeling of microbial stains *in silico*," *Trends in Biochemical Sciences*, **26**: 179-186 (2001); Edwards J., and Palsson, B, Metabolic flux balance analysis and the *in silico* analysis of Escherichia coli K-12 gene deletions, *BMC Structural Biology*, 1(2) (2000a); Edwards J.S., and Palsson, B, O., Systemic properties of the *Haemophilus influenzae* Rd metabolic genotype, *Journal of Biological Chemistry*, 274(25):17410-16, (1999); Karp P. D. et al., HinCyc: A knowledge base of the complete genome and metabolic pathways of H. influenzae, *ISMB* 4:116-24, (1996); Karp P. D. et al., The EcoCyc and MetaCyc databases, *Nucleic. Acids Res.* 28(1):56-59 (2000); Ogata et al., KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* 27(1):29-34 (1999); Schilling C. H. and Palsson B. O., Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis, *J. Theor. Biol.*, 203(3): 249-83 (2000); Selkov E. Jr. et al., MPW: the metabolic pathways database, *Nucleic Acids Res.*, 26(1): 43-45 (1998); Selkov E. et al., A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* 197(1-2):GC11-26 (1997)). This process involves the use of annotated genome sequences, and biochemical and physiological data. These annotated genome sequences and biochemical and physiological data can be found in one or more internal or external databases, such as those described in detail in the discussion of the computer systems of the current invention below. Careful analysis of the reconstructed network is needed to reconcile all the data sources used. Similar methods can be used for the reconstruction of other biochemical reaction networks.

[0049] A method of this aspect of the present invention then uses the reconstructed comprehensive, substantially whole, or whole biochemical reaction network to determine optimal properties of the comprehensive, substantially whole, or whole biochemical reaction network under specified and varying environmental conditions. This determination allows the design of a biochemical reaction network that achieves a desired performance in a specified environment. This in turn, can be combined with steps for constructing the genetic makeup of a cell and cultivating the cell, described below, to provide a method for developing a recombinant cell, or a population of cells, that achieves the desired performance.

[0050] Optimal properties of the comprehensive, substantially whole, or whole biochemical reaction network under a series of specified environments can be determined using computational methods known as optimization methods. Optimization methods are known in the art (*see e.g.,* Edwards and Palsson (1999)). The optimization methods used in the methods of the current invention can, for example and not intended to be limiting, utilize a combination of flux balance analysis (FBA), phase plane analysis (PhPP), and a determination of a Line of Optimality (LO), as described in further detail below.

[0051] The reconstructed metabolic network can then be used to perform quantitative simulations of the metabolic flux distribution in a steady state using established methods (Bonarius et al., Flux analysis of underdetermined metabolic networks: The quest for the missing constraints, *Trends in Biotechnology*, 15(8): 308-14 (1997); Edwards J.S., et al., Metabolic flux Balance Analysis, In: (Lee S. Y., Papoutsakis E.T., eds.) Metabolic Engineering: Marcel Deker. P 13-57 (1999); Varma A. and Palsson B.0, Metabolic flux balancing: Basic concepts, Scientific and practical use, *Bio/Technology* 12:994-98 (1994a)). Computer simulations of the metabolic network can be performed under any conditions. Furthermore, any reaction list can be simulated in a computer by changing the parameters describing the environment and the contents of the reaction list.

[0052] The metabolic capabilities of a reconstructed metabolic network can be assessed using the established method of flux balance analysis (FBA) (Bonarius et al., (1997); Edwards et al. (1999); Varma and Palsson (1994a)). FBA is based on the conservation of mass in the metabolic network in a steady state and capacity constraints (maximal fluxes through the reactions) on individual reactions in the network. Additionally, experimentally determined strain specific parameters are also required, the biomass composition (Pramanik J. and Keasling J. D., Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements, *Biotechnology and Bioengineering*, 56(4): 398-421 (1997)) and the maintenance requirements (Varma A. and Palsson B. O., Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology*, 60(10): 3724-31 (1994b)). These factors are then used to calculate the flux distribution through the reconstructed metabolic network.

[0053] More specifically, the definition of these factors leads mathematically to a closed solution space to the equations in which all feasible solutions lie. There are thus many possible solutions (flux distributions) to the problem. The 'best' or optimal solution within the set of all allowable solutions can then be determined using optimization procedures and a stated objective. The optimization procedure used has been linear programming and the objective is the optimal use of the biochemical reaction network to produce all biomass components simultaneously. These optimization procedures are established and have been published (Varma and Palsson (1994a); Bonarious (1997); and Edwards et al. (1999)). The comparison of the calculated behavior based on the optimal growth objective to the experimental data is favorable in the majority of cases (Varma (1994b); Edwards J.S., Ibarra R.U., and Palsson B.O.(herein incorporated by reference), In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data, *Nat Biotechnol.*, 19(2): 125-30 (2001a); and Edwards, Ramakrishna, and Palsson, Characterizing phenotypic plasticity: A phenotype phase plane analysis, *Biotech Bioeng*, In Press, (2001b)). In other words, these solution confinement and optimization procedures lead to a prediction of the optimal uses of a biochemical reaction network to support cellular growth and for pre-evolved strains give a good estimate of actual biological function.

[0054] The use of alternate survival objectives, such as sporulation, and scouring of rare elements under nutritionally poor conditions, has not been described. Competition and evolution under these conditions can also be used to define and generate optimal network functions.

[0055] These procedures lead to the calculation of optimal function under a single growth condition. This is very limiting and a method to analyze a large number of growth conditions is needed.

[0056] As stated above, all steady state metabolic flux distributions are mathematically confined to the solution space defined for a given reconstructed metabolic network, where each solution in the solution space corresponds to a particular flux distribution through the network or a particular metabolic phenotype (Edwards and Palsson (1999)). Under a single specified growth condition, the optimal metabolic flux distribution in the cone can be determined using linear programming (LP) or other related approaches for calculating optimal solutions of such

problems. If the constraints vary, the shape of the cone changes and the optimal flux vector may qualitatively change. Phenotype Phase Plane (PhPP) analysis considers all possible variations in two or more constraining environmental variables. This method is now disclosed.

[0057] Uptake rates of two nutrients (such as the carbon substrate and oxygen) can be defined as two axes on an (x,y)-plane, and the optimal flux distribution can be calculated for all points in this plane using the procedures described above. When this procedure is implemented for a reconstructed metabolic network that is biologically realistic, we find that there are a finite number of qualitatively different optimal metabolic flux maps, or metabolic phenotypes, present in such a plane. The demarcations on the phase plane can be defined by using shadow prices of linear optimization (Chvatal V., Linear Programming, (New York: W. H. Freeman and Co.) (1983)). The procedure described leads to the definition of distinct regions, or "phases", in the (x,y)-plane, for which the optimal use of the metabolic network is qualitatively different. Each phase can be designated as Pnx,y, where P represents a particular phenotype or flux distribution, n is the number arbitrarily assigned to the demarcated region for this phenotype, and the two uptake rates form the axis of the plane.

[0058] The phase planes can be constructed by calculating the shadow prices throughout the two-parameter space, and lines are drawn to demarcate regions of constant shadow prices. The shadow prices define the intrinsic value of each metabolite toward the objective function (Chvatal (1983)). The shadow prices are either negative, zero, or positive, depending on the value of the metabolite to optimizing growth under a given environmental condition, as represented by particular numerical values of the uptake rates represented by the x and y axes. When shadow prices become zero as the values of the uptake rates are changed there is a qualitative shift in the optimal metabolic map. This criterion defines the lines in the PhPP.

[0059] The line of optimality (LO) is defined as the line representing the optimal relation between the two uptake fluxes corresponding to the axes of the PhPP. For aerobics, this line can be interpreted as the optimal oxygen uptake for the complete oxidation of the substrate in order to support the maximal biomass yield.

[0060] The metabolic reconstruction and phenotype phase plane analysis procedures are then used to predict the conditions under which the desired metabolic behavior, for example

maximum uptake rates, will be optimal. In other words, metabolic reconstruction and PhPP are used to determine optimal performance. The maximal uptake rates lead to the definition of a finite rectangular region in the phase plane. The optimal growth condition within this rectangle will then be the predicted outcome of an evolutionary process within the given constraints.

[0061] Using the optimization procedure, the properties of the corresponding actual biochemical reaction network may not be optimal or the same as desired from a practical standpoint. The simulated reconstructed network and its synthesis in an organism may not display the optimal solution desired, also referred to herein as the desired optimal performance or desired optimal function. Lack of optimality may be due to the fact that:

[0062] 1. The natural organism with an intact network has never experienced the environmental conditions of interest and never undergone growth competition and selection in this environment, or

[0063] 2. The man made network is perturbed from its evolutionarily determined optimal state by genetic manipulations, through the deletion/addition of a new reaction from/to the network.

[0064] The *in silico* methods of the current invention are designed to resolve this second cause of lack of optimality, by altering the reactions in the network until a desired performance is achieved. Then culturing methods, which can be included in the method of the current invention as described in further detail below, can be used to resolve the first cause of the lack of optimality related to growth competition and selection.

[0065] As mentioned above, after calculation of the optimal properties, a metabolic engineer can alter the reaction list in the network, or an algorithm can be developed that automatically alters one or more reactions in the reaction list, to achieve a desired performance. After alteration of the biochemical list, optimal properties of this network under given environmental conditions can be calculated. This is an iterative design procedure that may require many different versions of the reaction list until the desired performance is achieved. The desired performance is a qualitative characteristic or quantitative value for a property calculated using an optimization procedure. Many properties for which a desired performance can be achieved are

known in the art. For example, a desired performance can be a desired growth rate or a desired yield of a biomolecule such as an enzyme or an antibiotic.

[0066] The optimization method may be carried out using a computer system provided by the present invention. The computer system typically includes a database that provides information regarding one or more biochemical reaction networks of at least one organism; a user interface capable of receiving a selection of one ~~two~~ or more biochemical reaction networks for optimization and/or comparison, and capable of receiving a selection of a desired performance; and a function for carrying out the optimization method calculations and recalculations. Furthermore, the computer system of the present invention may include a function for performing biochemical reaction network reconstruction described hereinabove.

[0067] The computer system can be a stand-alone computer or a conventional network system including a client/server environment and one or more database servers. A number of conventional network systems , including a local area network (LAN) or a wide area network (WAN), are known in the art. Additionally, client/server environments, database servers, and networks are well documented in the technical, trade, and patent literature. For example, the database server can run on an operating system such as UNIX, running a relational database management system, a World Wide Web application, and a World Wide Web Server.

[0068] Typically, the database of the computer system of the present invention includes information regarding biochemical reactions in a comprehensive biochemical reaction network, a substantially whole biochemical reaction network, or a whole biochemical reaction network. This information can include identification of biomolecular reactants, products, cofactors, enzymes, rates of reactions coenzymes, etc. involved in at least some of the biochemical reactions of the network. This information can include the stoichiometric coefficients that indicate the number of molecules of a compound that participates in a particular biochemical reaction. This information can include any and all isozymes that are found in the organism. This information can include all the related biochemical reactions that can be catalyzed by a single enzyme. This information can include the formation of oligomeric enzyme complexes, that is when many different protein molecules must non-covalently associate to form a complex that can carry out the chemical reactions. This information can include the location of the enzyme that

carries out the reaction, (i.e. if it is in a membrane, in the cytoplasm, or inside an organelle such as the mitochondria). The information can also include experimentally derived or calculated rates of reactions under various conditions, biomass compositions, and maintenance requirements. This information can include non-mechanistic growth and non-growth associated maintenance requirements. This information can include mechanistic maintenance information such as inefficiency in protein synthesis. This information can include data derived from genome-scale mRNA or protein expression profiles. This information can include data on operon or regulatory structures that are associated with the expression of a particular gene. This information can include sequence variations that reflect changes in enzyme properties. This information can include a condition-dependent inclusion of a biochemical reaction, depending for instance if a gene is not expressed under the conditions of interest. Where the biochemical reaction network is a metabolic reaction network, the information for example can include known consumption rates, by-product production rates, and uptake rates.

[0069] The information in the database may include biomolecular sequence information regarding biomolecules involved in the biochemical reactions of the biochemical reaction network, for example information regarding multiple biomolecular sequences such as genomic sequences. At least some of the genomic sequences can represent open reading frames located along one or more contiguous sequences on each of the genomes of the one or more organisms. The information regarding biochemical reaction networks can include information identifying those biochemical reaction networks to which a biomolecular sequence plays a role and the specific reactions in the biochemical reaction network involving the biomolecule.

[0070] The database can include any type of biological sequence information that pertains to biochemical reactions. For example, the database can be a nucleic acid sequence database, including ESTs and/or more preferably full-length sequences, or an amino acid sequence database. The database preferably provides information about a comprehensive, substantially whole, or whole biochemical reaction network. For example, the database can provide information regarding a whole metabolic reaction network. The database can provide nucleic acid and/or amino acid sequences of an entire genome of an organism.

[0071] The database can include biochemical and sequence information from any living organism and can be divided into two parts, one for storing sequences and the other for storing information regarding the sequences. For example, the database can provide biochemical reaction information and sequence information for animals (e.g., human, primate, rodent, amphibian, insect, etc), plants, or microbes. The database is preferably annotated, such as with information regarding the function, especially the biochemical function, of the biomolecules of the database. The annotations can include information obtained from published reports studying the biochemistry of the biomolecules of the database, such as specific reactions to which a biomolecule is involved, whether the biomolecule is or encodes an enzyme, whether the sequence is a wild-type sequence, etc.

[0072] The annotations and sequences of the database provide sufficient information for a selected biochemical genotype of an organism to be identified. A biochemical genotype is a grouping of all the nucleic acid or amino acid sequences in a selected biochemical process of an organism. For example, a metabolic genotype is a grouping of all the nucleic acid and/or amino acid sequences of proteins involved in metabolism. Methods for identifying metabolic genotypes have been described in the literature (*see e.g.* Edwards and Palsson 1999).

[0073] The database can be a flat file database or a relational database. The database can be an internal database, or an external database that is accessible to users, for example a public biological sequence database, such as Genbank or Genpept. An internal database is a database maintained as a private database, typically maintained behind a firewall, by an enterprise. An external database is located outside an internal database, and is typically maintained by a different entity than an internal database. A number of external public biological sequence databases are available and can be used with the current invention. For example, many of the biological sequence databases available from the National Center for Biological Information (NCBI), part of the National Library of Medicine, can be used with the current invention. Other examples of external databases include the Blocks database maintained by the Fred Hutchinson Cancer Research Center in Seattle, and the Swiss-Prot site maintained by the University of Geneva. Additionally, the external databases can include a database providing information regarding biochemical reactions, including databases of published literature references describing and analyzing biochemical reactions. Where a database included in the computer systems of the

present invention is a public computer database that does not identify information that is relevant for a particular biochemical reaction network, the computer system either includes a function for performing biochemical reaction network reconstruction, or includes identification of the database entries that pertain to a particular biochemical reaction network. Additionally, there are several databases with biochemical pathway information, these databases include, for non-limiting example, EcoCyc, KEGG, WIT, and EMP. These databases can be used to provide the information to reconstruct the metabolic models.

[0074] In addition to the database discussed above, the computer system of the present invention includes a user interface capable of receiving a selection of one or more biochemical reaction networks for optimization and/or comparison, and capable of receiving a selection of an optimal performance. The interface can be a graphic user interface where selections are made using a series of menus, dialog boxes, and/or selectable buttons, for example. The interface typically takes a user through a series of screens beginning with a main screen. The user interface can include links that a user may select to access additional information relating to a biochemical reaction network.

[0075] The function of the computer system of the present invention that carries out the optimization methods typically includes a processing unit that executes a computer program product, itself representing another aspect of the invention, that includes a computer-readable program code embodied on a computer-usable medium and present in a memory function connected to the processing unit. The memory function can be, for example, a disk drive, Random Access Memory, Read Only Memory, or Flash Memory.

[0076] The computer program product that is read and executed by the processing unit of the computer system of the present invention, includes a computer-readable program code embodied on a computer-usable medium. The program code is capable of interacting with the database and effects the following steps within the computing system: providing an interface for receiving a selection of a desired performance of the networks; determining the desired optimal properties, displaying the results of the determination, and altering the biochemical reaction network, before recalculating optimal properties of the biochemical reaction network, and repeating the process until a desired performance is achieved. Altering the biochemical reaction network can be

performed based on an alteration identified by a user, or can be performed automatically by the program code. The computer program can further provide an identification of database entries that are part of a reconstructed biochemical network, or can perform biochemical reaction network reconstruction. Furthermore, the computer program of the present invention can provide a function for comparing biochemical reaction networks to identify differences in components and properties.

[0077] The computer-readable program code can be generated using any well-known compiler that is compatible with a programming language used to write software for carrying out the methods of the current invention. Many programming languages are known that can be used to write software to perform the methods of the current invention.

[0078] As mentioned above, a method of this aspect of the invention can further include steps that involve adaptive evolution of a cultured strain to achieve the desired performance. Virtually any cell can be used with the methods of the present invention including, for example, a prokaryotic cell, or a eukaryotic cell such as a fungal cell or an animal cell including a cell of an animal cell line. However, a biochemical reaction network of the cell, or the cell of a closely related organism, must be sufficiently characterized to allow a high quality reconstruction of the comprehensive, substantially whole, and/or whole biochemical reaction network in a computer. Preferably, essentially the entire genome of the organism has been sequenced and genes encoding biomolecules, typically proteins, involved in the biochemical reaction network have been identified.

[0079] The genetic makeup of a cell can be constructed to contain the biochemical reactions that meet the desired performance to produce a cell with a potential to meet the desired performance. This can be achieved using the indigenous list of reactions in the cell and by adding and subtracting reactions from this list using genetic manipulations to achieve the reaction list capable of achieving the desired performance criteria, identified by the steps *performed in silico* described above. For example, reactions can be added or subtracted from the list by adding, changing, or deleting all or portions of one or more genes encoding one or more biomolecules involved in the reaction, for example by adding, changing, or deleting protein coding regions of one or more genes or by adding, changing, or deleting regulatory regions of

one or more genes. In addition, for example, reactions can be added or subtracted from the list by altering expression of regulatory components (e.g., transcription factors) that effect the expression of one or more biomolecules involved in one or more reactions of the reaction list. The resulting engineered cell may or may not display the optimal properties calculated ahead of time by the *in silico* methods using the iterative optimization procedure described above.

[0080] Many methods exist in the art that describe the genetic manipulations of cells (*see e.g.*, Datsenko K. A. and Wanner B. L., One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products, *Proc. Natl. Acad. Sci. U.S.A.*, 97(12):6640-45 (2000); Link et al., Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: Application to open reading frame characterization, *J. of Bacteriology*, 179:6228-37 (1997)). Any of the existing genetic manipulation procedures can be used to practice the invention disclosed.

[0081] After the cell has been constructed to have a potential to meet the desired performance, it is placed in culture under a specified environment. The specified environment is determined during the optimization procedure. That is, the optimization procedure calculates properties of the network under various environments, as described above, and identifies the specified environment in which the desired performance is achieved.

[0082] The cells are cultured for a sufficient period of time and under conditions to allow the cells to evolve to the desired performance. That is, adaptive evolution of natural or engineered strains is carried out as guided by the general optimization methods or procedures. Natural strains that have not experienced a particular environment or genetically altered strains can be analyzed by the network reconstruction and optimization procedures disclosed above. These strains can then be put under a selection pressure consistent with the desired function of the organisms and evolved towards the predetermined performance characteristics. The cells may achieve the desired performance without additional adaptive evolution. That is, the sufficient period of time for culturing the cells, may be immediately after the genetic makeup of the cells is constructed using the methods of the present invention without the need for further adaptive evolution.

[0083] In other words, Extended cultivation of a non-optimal or non-evolved strain can be performed to optimize or evolve the metabolic network toward the optimal solution that is achievable under the defined environmental conditions. The practice of this evolutionary process requires on the order of weeks to years to optimize a metabolic network depending on how far it is from the optimal conditions at the beginning of the evolutionary process, and how difficult it is to achieve the necessary changes through random mutation and shifts in regulation of gene expression. This process can be accelerated by the use of chemical mutagens and/or radiation. Additionally, the process is accelerated by genetically altering the living cell so that it contains the biochemical reactants determined by the *in silico* method described above, that achieve a desired performance.

[0084] Methods are known in the art for culturing cells under specified environmental conditions. For example, if the cell is *E. coli*, and the desired performance is a desired growth rate, the procedure set out below can be used. This procedure can be readily adapted for use with other bacterial cells and/or other performance criteria as is known in the art. Additionally, the procedures can be readily developed for use with other cell types such as animal cells. For example, the methods can be readily adapted for use with other culturing systems, such as large scales systems in which cells adhere to a culturing vessel. The culturing methods may be adapted for high-throughput analysis, as known in the art.

[0085] If a strain needs to be directionally evolved to achieve the desired performance, then following the construction of the metabolic reaction network in the chosen host strain, the cells are typically stored frozen at –80°C with 30% glycerol. For each adaptive evolutionary process, frozen stocks are plated on LB agar and grown overnight at 37°C. From the plate, individual colonies can be identified that arose from a single cell. An individual colony can be used to inoculate a liquid culture, known as a pre-culture. Pre-cultures inoculated from a single colony of the respective strain are grown overnight in the defined medium for the subsequent evolutionary process. A pre-culture sample is taken the following day, typically at mid-log phase (in the middle of logarithmic growth) of growth to inoculate the culture conditions that define the environment that the adaptive evolution is to take place. Batch bioreactors or other suitable culture vessels are then initiated. This, typically would be done at 250mL volumes in micro-carrier spinner flasks inside a temperature controlled incubator on top of a magnetic stir plate, set

at suitable --typically high-- speed to ensure sufficient aeration and at the optimal growth temperature (37°C for wildtype *E. coli*) for any given strain. Other frequently used cultivation procedures known in the art can also be used.

[0086] After a suitable time period, typically the following day for *E. coli* (before the culture reaches stationary phase), an aliquot of the culture now in mid-log phase is serially transferred to a new spinner flask containing fresh medium. If the culture is being optimized for growth rate, stationary phase must be avoided to ensure that the selection criterion is growth rate. Then serial transfers are performed at fixed time intervals (typically every 24 hours depending on the growth rate) at mid-log phase and the volume of the innoculum into the new culture vessel is adjusted accordingly based on the increase in growth rate.

[0087] Growth rate is thus monitored frequently, typically on a daily basis in order to determine the proper volume of the inoculum to use for the next serial transfer. This serial cultivation process is repeated sufficiently often to allow the cells to evolve towards its optimal achievable growth under the conditions specified through the medium composition.

[0088] The growth and metabolic behavior is monitored during the adaptive evolutionary process to determine how the population is evolving over time. At fixed time intervals typically every few days, the culture is tested for metabolic and growth behavior, by measuring the oxygen uptake rate, substrate uptake rate and the growth rate. The results are then plotted as a data point on the phenotype phase plane. Movement of the so determined data point towards the line of optimality would indicate evolution towards optimal growth behavior. These measurements of the membrane transport fluxes along with the growth rate are repeated until we observe that the cells are operating their metabolic network such that the data points lie at the optimal conditions. The evolutionary process can then be continued until there is no further increase in the optimal performance, i.e., growth rate. If no further change is observed then we have achieved the maximal growth rate for the given conditions.

[0089] Byproduct secretion can be monitored by HPLC or other suitable methods of analytical chemistry to assess changes in metabolism that are implicated in the evolution towards optimal growth behavior. For these studies it is also imperative to determine a correlation of dry weight vs optical density for the evolved strain since this will be different from the wildtype. In

addition to monitoring the growth rate and steady state growth, the cultures are inspected for any signs of possible contamination or co-evolution with a mutant subpopulation-aliquots for each day of culture are kept refrigerated as a backup in the event of any contamination, and the phenotype of the culture is ascertained by plating samples of the culture and inspecting for any differences in colony morphology or different mutants. On a daily basis, the optical density of the culture, time of inoculation, innoculum volume, growth rate, and any signs of contamination, are logged. Samples are also frozen at – 80 °C in 30% glycerol for each day of culture for any possible further use.

[0090] After cells produced using the methods of the current invention evolve to achieve optimal performance, they may be further characterized. For example, a characterization can be made of the biochemical reaction network(s) of the cell. This characterization can be used to compare the properties, including components, of the biochemical reaction network(s) in the living cells to those predicted using the *in silico* methods.

[0091] The following examples are intended to illustrate but not limit the invention.

EXAMPLE 1

DETAILED METHODS FOR DETERMINATION OF OPTIMAL FUNCTION AND EVOLUTION FOR *E.COLI*

[0092] This example provides cultivation procedures that can be used to determine the optimality of strain performance and to carry out adaptive evolutionary processes.

[0093] Growth behavior of *E. coli* is determined by the following standard procedures. Growth is carried out in M9 minimal media (Maniatis et al., Molecular Cloning: A Laboratory Manual, (Cold Springs Harbor, N.Y., Cold Spring Harbor Laboratory, 545 (1982)) with the addition of the carbon source (Table 1). Cellular growth rate is varied by changing the environmental conditions, i.e. by changing the carbon source concentration approximately

ranging between 0.05 – 4 g/L, temperature (27.5 °C to 37 °C), and oxygen (0-100% saturation relative to air). Batch cultures are set up in bioreactors at volumes of 250 mL in 500 mL flask with aeration. For these cultures the oxygen uptake rate (OUR) is monitored online, by either measuring the mass transfer coefficient for oxygen (k_la), by using an off-gas analyzer, or by monitoring the oxygen tension in a respirometer chamber using known methods in the art. The temperature is controlled using a circulating water bath (Haake, Berlin, Germany). All measurements and data analysis are restricted to the exponential phase of growth. The biomass and the concentration of the substrate and metabolic by-products in the media are monitored throughout the experiment using methods known in the art. Cellular growth is monitored by measuring the optical density (OD) at 600 nm and 420 nm and by cell counts. OD to cellular dry weight correlations are determined by two different methods; (1) 50 mL samples of culture are spun down and are dried at 75°C to a constant weight, and (2) 25 ml (taken throughout the culture) samples are filtered through a 0.45um filter and dried to a constant weight. The concentration of metabolites in the culture media is determined by HPLC. An aminex HPX-87H ion exchange carbohydrate-organic acid column (@ 66°C) is used with degassed 5 mM H_2SO_4 as the mobile phase and UV detection. Enzymatic assays are also used to determine the substrate uptake rate and by-product secretion rates. The dissolved oxygen in the culture is monitored using a polarographic dissolved oxygen probe. Oxygen consumption is measured by three different methodologies; (1) passing the effluent gas through a 1440C Servomex oxygen analyzer (Servomex Co., Inc. Norwood, MA), (2) calculated from the dissolved oxygen reading and k_la measurements, and (3) in a respirometer chamber in a separate 50 ml flask. All three methods used for measuring the oxygen uptake rate give similar and reproducible results.

[0094] The cultivation procedures provided in this Example can be used to determine the optimality of strain performance and to carry out adaptive evolution.

Table I.

M9 Minimal Media (Per liter)		
5 X M9 Salts		200mL
1M $MgSO_4$		2mL
20% Solution of Carbon Source		20 mL
1M CaCl		0.1mL
5 X M9 Salts (Per liter)		
$Na_2HPO_4*7H_20$		64g
KH_2PO_4		15g
NaCl		2.5g
NH_4Cl		5.0g

EXAMPLE 2

CALCUATION OF OPTIMALITY PROPERTIES AND PHENOTYPIC PHASE PLANES

[0095] This example shows how we calculate the optimality properties of the reconstructed network and how such results are represented on a phenotypic phase plane.

[0096] The capabilities of a metabolic network can be assessed using flux balance analysis (FBA) (Bonarius et al., (1997); Edwards et al., (1999); Varma et al., (1994a); Varma et al., (1994b)). FBA is primarily based on the conservation of mass in the metabolic network. The conservation requirement is implemented by stoichiometric balance equations; thus, FBA relies on the stoichiometric characteristics of the metabolic network.

[0097] The flux balance equation is $S \bullet v = b^v$, where S is the stoichiometric matrix, the vector v defines the metabolic fluxes, and b^v is nominally zero – thus, enforcing simultaneous mass, energy, and redox balance constraints through a set of mass balances. Variations of the b^v vector from zero were used in the shadow price analysis (discussed below). In the *E. coli* metabolic network, the number of metabolic fluxes was greater than the number of mass balances, thus leading to a plurality of feasible flux distributions that lie in the null space of the matrix S. Additional constraints were also placed on the feasible value of each flux in the metabolic

network ($\alpha_i \leq v_i \leq \beta_i$). These constraints were utilized to define the reversibility of the internal reactions and to set the uptake rate for the transport reactions. The transport of inorganic phosphate, ammonia, carbon dioxide, sulfate, potassium, and sodium were unrestrained; whereas the uptake of the carbon source and oxygen were restrained as specified. All metabolic by-products (i.e. acetate, ethanol, formate, pyruvate, succinate, lactate, etc) which are known be transported out of the cell were always allowed to leave the metabolic system. In this analysis, α_i for the internal fluxes was set to zero for all irreversible fluxes and all reversible fluxes were unbounded in the forward and reverse direction (the reversibility of each reaction is defined on the website of supplementary information). The intersection of the null space, and region defined by the linear inequalities, defined the capabilities of the metabolic network and has been referred to as the flux cone (Schilling et al., 1999).

[0098] The determination of the optimal metabolic flux distribution that lies within the flux cone was formulated as a linear programming (LP) problem, in which the solution that minimizes a particular metabolic objective was identified (Bonarius, H. P .J. et al., Metabolic flux analysis of hybridoma cells in different culture media using mass balances, *Biotechnology and Bioengineering* 50: 299-318 (1996);Edwards et al., (1999);Pramanik et al., (1997); Varma et al., (1994a); Varma et al., (1994b)). The growth flux was defined as the objective. The growth flux was defined as a metabolic flux utilizing the biosynthetic precursors, X_m, in the appropriate ratios, $\sum_{all\ m} d_m \cdot X_m \xrightarrow{v_{growth}} Biomass$, where d_m is the biomass fraction of metabolite X_m. The biomass composition was defined based on the literature (Neidhardt, FC, Ed., *Escherichia coli* and *Salmonella*: cellular and molecular biology (ASM Press, Washington, D.C., 1996); Neidhardt, FC, Umbarger, HE, *in Escherichia coli* and *Salmonella* : cellular and molecular biology F. C. Neidhardt, Ed. (ASM Press, Washington, D.C., 1996), vol. 1, pp. 13-16 (1996)).

[0099] All steady state metabolic flux distributions are mathematically confined to the flux cone defined for the given metabolic map, where each solution in the flux cone corresponds to a particular internal flux distribution or a particular metabolic phenotype (Varma et al., (1994a); Varma et al., (1994b)). Under specified growth conditions, the optimal phenotype in the cone can be determined using linear programming (LP). If the constraints vary, the shape of the cone changes and the optimal flux vector may qualitatively change; for example, inactive fluxes may

be activated and vice versa. The phase plane analysis is developed to consider all possible variations in two constraining environmental variables.

[0100] Defining Phenotype Phase Planes (PhPPs): Uptake rates of two nutrients (such as the carbon substrate and oxygen) were defined as two axes on an (x,y)-plane, and the optimal flux distribution was calculated for all points in this plane. There are a finite number of qualitatively different optimal metabolic flux maps, or metabolic phenotypes, present in such a plane. The demarcations on the phase plane were defined by a shadow price analysis (Varma, A, Boesch, BW, Palsson, BO., Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates, *Applied and Environmental Microbiology* 59, 2465-73 (1993); Varma, A, Palsson, BO., Metabolic capabilities of Escherichia coli: II. Optimal growth patterns. *Journal of Theoretical Biology* 165, 503-522 (1993)). This procedure led to the definition of distinct regions, or "phases", in the plane, for which the optimal use of the metabolic pathways was qualitatively different. Each phase was written as $Pn_{x,y}$, where P represents phenotype, n is the number of the demarcated region for this phenotype (as shown in the corresponding Fig. 1), and x,y the two uptake rates on the axis of the plane.

[0101] Calculating the Phase Plane: The phase planes were constructed by calculating the shadow prices throughout the two-parameter space, and lines were drawn to demarcate regions of constant shadow prices. The shadow prices defined the intrinsic value of each metabolite toward the objective function. Changes in shadow prices were used to interpret of metabolic behavior.

[0102] Mathematically, the shadow prices are defined as,

[0103] $$\gamma_i = -\frac{dZ}{db^v_i} \qquad (1)$$

[0104] and are associated with each metabolite in the network. The shadow prices defined the sensitivity of the objective function (Z) to changes in the availability of each metabolite (b^v_i defines the violation of a mass balance constraint and is equivalent to an uptake reaction). The shadow prices were either negative, zero, or positive, depending on the value of the metabolite. The direction and magnitude of the shadow price vector in each region of the phase plane was

different (by definition of the phase plane); thus, the state of the metabolic system was different in each region.

[0105] Isoclines: Isoclines were also defined to interpret the metabolic phenotype. Isoclines were defined to represent the locus of points within the two-dimensional space that provide for the same value of the objective function. The slope of the isoclines within each region was calculated from the shadow prices; thus, by definition the slope of the isoclines was different in each region of the PhPP. A ratio of shadow prices was used to define the slope of the isoclines (ρ):

[0106]
$$\rho = -\frac{\gamma_x}{\gamma_y}\bigg|_Z = -\left(\frac{-dZ/db^v{}_x}{-dZ/db^v{}_y}\right)\bigg|_Z = -\frac{db^v{}_y}{db^v{}_x}\bigg|_Z \qquad (2)$$

[0107] The negative sign in Eqn. 2 was introduced in anticipation of its interpretation. The condition dependent relative value of the substrates, defined as ρ, was used to interpret the constraining factors on the metabolic network. In regions where ρ was negative, there was dual limitation of the substrates. Under different condition, the isoclines were also either horizontal or vertical in certain phase plane regions, representing regions of single substrate limitation, the ρ value in these regions was zero or infinite, respectively. Certain regions in the PhPP also had a positive ρ; these regions were termed "futile" regions, and increased uptake of one of the substrates had a negative effect on the objective function in these regions. Finally, due to stoichiometric limitations, there were infeasible steady state regions in the PhPP.

[0108] Line of Optimality: The line of optimality (LO) was defined as the line representing the optimal relation between the two metabolic fluxes corresponding to the axis of the PhPP. For results presented herein, this line can be interpreted as the optimal oxygen uptake for the complete oxidation of the substrate in order to support the maximal biomass yield.

[0109] These procedures were applied to the reaction list for *E. coli* K-12 defined by (Edwards and Palsson, The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities, *Proc. Natl. Acad. Sci. U.S.A.*, 97(10):5528-33 (2000b)). It generates the two and three dimensional phase planes that are used in the examples below.

EXAMPLE 3

OPTIMAL BEHAVIOR OF *E. COLI* UNDER DEFINED CONDITIONS

[0110] This Example shows that the strain used exhibited optimal aerobic growth using acetate and succinate as primary substrates without adaptive evolution.

[0111] The list of metabolic reactions that take place in *E. coli* K-12 M1655 has been assembled (Edwards and Palsson (2000b)). Based on this list a stoichiometric matrix was formulated. Using maximal uptake rates for oxygen (on the y-axis) and a carbon substrate (on the x-axis) a phenotypic phase plane was calculated using the procedures described above. Specifically two carbon sources were used, acetate and succinate. Then the calculated phase planes were used to determine the optimal growth conditions and a series of growth experiments were performed. The computational (i.e. *in silico*) and experimental results were then compared.

[0112] **Acetate.** Optimal growth performance on acetate was investigated in silico, and the predictions generated were compared to experimental data. The in silico study started with a phenotype phase plane (PhPP) analysis with the acetate and oxygen uptake rates defined as the axes of the two-dimensional projection of the flux cone representing the capabilities of *E. coli* metabolism (Fig. 1). The flux cone is the region of all admissible steady state metabolic flux distributions (for a complete description of the flux cone see ref (Schilling et al., Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era, *Biotechnol. Prog.*, 15(3):296 (1999)). Furthermore, a three-dimensional projection of the flux cone with the growth rate defined as the third dimension was utilized (Fig. 1). The in silico analysis of the acetate-oxygen PhPP has been described in Example 2 above. The optimal (with respect to cellular growth) relation between the acetate and oxygen uptake rate, and this line is referred to as the line of optimality (LO).

[0113] The PhPP was used to analyze and interpret the operation of the metabolic network. For example, under oxygen limitations the characteristics of the metabolic network may be defined by region 2 of the PhPP (Fig 1&2), where the acetate uptake rate exceeds the optimal relation to the oxygen uptake rate. From Fig. 1, it can be seen that if the metabolic network were

operating within region 2, the optimal capability to support growth would be increased by reducing the acetate uptake rate to a point on the LO. A similar interpretation can be made for points within region 1, with oxygen and acetate switching roles. Hence, metabolic flux vectors defining a point in region 1 or region 2 would indicate inefficient utilization of the available resources. Thus, the in silico PhPP analysis led to the conclusion that if the regulation of the *E. coli* metabolic network has evolved to operate optimally to support growth with acetate as the sole carbon source, the relation between the acetate and oxygen uptake rate and the growth rate should be defined by the LO (Fig. 1&2).

[0114] The relation between the acetate and oxygen uptake rates and the growth rate was experimentally examined by cultivating *E. coli* MG1655 on acetate minimal medium. The acetate uptake rate was experimentally controlled by changing the acetate concentration in the minimal media. The uptake rates of acetate and oxygen and the growth rate were measured and the experimental points were plotted on the PhPP (Fig. 1 and 2). The calculated optimal relation between the acetate and oxygen uptake rate was then compared to the experimental data (Fig 1). Comparison of the experimental data to the in silico predictions indicated a 14% difference between the slope (0.91) of the linear regression line for the experimental data and the slope (1.04) of the in silico defined LO for aerobic growth on acetate minimal media.

[0115] The measured and calculated growth rates were plotted as the third-dimension above the acetate-oxygen PhPP (Fig. 2). The color-coded surface represents the three-dimensional projection of the flux cone. In other words, the color-coded surface defines the solution space, and all feasible steady state metabolic flux distributions are confined within the surface. Yes we can. The LO on the two-dimensional phase plane (Fig. 1) is a projection of the edge on the three-dimensional surface on to the x,y-plane (acetate uptake rate, oxygen uptake rate). The experimental data were plotted in the three-dimensional space (Fig. 2). To quantitatively visualize the proximity of the data points to the LO in three-dimensions, the in silico predictions and the experimental data were projected onto each plane formed by the basis vectors.

[0116] The projection of the three-dimensional LO and the experimental data points onto the (x,y), (x,z), and (y,z) (x-axis: acetate uptake rate; y-axis: oxygen uptake rate; z-axis: growth rate) planes is indicated in Figures. 3&4 respectively, where the quality of the linear regression is

indicated by the correlation coefficient, and the data are compared to the in silico predictions. The predicted and the observed metabolic fluxes (substrate and oxygen uptake rates and growth rate) for each point were directly compared and the in silico predictions and had an overall average error of 5.8%. At this point, we should note that the information used to reconstruct the metabolic network was obtained independent from the present experiments (Edwards and Palsson, (2000b)). The calculated PhPP represents a priori interpretation and prediction of the data obtained in the present study.

[0117] **Succinate**. The succinate-oxygen PhPP (Fig. 5) was more complicated than the acetate-oxygen PhPP. The succinate-oxygen PhPP (Fig. 5) had 4 distinct regions of qualitatively distinct optimal metabolic network utilization. Regions 1 and 4 of the succinate-oxygen PhPP were analogous to regions 1 and 2 of the acetate-oxygen PhPP. For example, it can be seen from Figure 5 that the maximal growth flux for a flux vector in region 4 can be increased if the succinate uptake is reduced to a point defined by the region 3,4 demarcation. Furthermore, from the PhPP analysis, region 3 is defined as a single substrate limited region. The single substrate limited region indicates that the succinate uptake rate has little effect on the maximal growth flux in region 3, whereas the oxygen uptake rate has a positive effect on the growth rate.

[0118] Region 2 is defined as a dual substrate limited region, since in region 2 the metabolic network can support an increased growth rate if the succinate uptake rate is increased. The in silico analysis shows that the cellular growth rate can be increased by operating the metabolic network off of the LO in region 2, by implementing a partially aerobic metabolism and the secretion of a metabolic by-product. The optimal metabolic by-product was calculated to be acetate. The production of a reduced metabolic by-product in region 2 however reduces the overall biomass yield. Therefore, based on the PhPP analysis, it may be surmised that, if the regulation of the metabolic network evolved toward optimal growth with succinate as the sole carbon source, the metabolic network will operate with a flux vector along the LO. However, the growth rate can be increased by moving the flux vector into region 2, thus we expect that the network should only operate in region 2 when oxygen is limited and succinate is plentiful if the stated hypothesis is true.

[0119] *E. coli* growth experiments on succinate minimal M9 media were performed to critically test the hypothesis given the above in silico analysis. Multiple batch cultures were grown at various succinate concentrations and temperatures to span a range of succinate uptake rates. The aeration and agitation were held constant to maintain a consistent maximal oxygen diffusion rate in all the cultures. The succinate and oxygen uptake rates and the growth rate were measured separately for each independent growth experiment. The experimental data were then directly compared to the in silico predictions (Fig. 5).

[0120] The experimental data points were consistent with the stated hypothesis: the flux vector consistently identified points along the LO for oxygen uptake rates below a critical value (~18.8 mmole·g-DW-1·hr-1). Furthermore, the cultures that identified points along the LO produced little or no acetate as a metabolic by-product (as predicted by the in silico analysis – see Fig. 5). As hypothesized, the experimental data indicates horizontal movement of the flux vector within region 2 for the experimental systems that are oxygen limited but have plentiful succinate. The break point in the experimental data was determined to correspond to a maximal oxygen uptake rate of 18.8 ± 0.5 mmole*gDW-1*hr-1. Flux vectors within regions 3 or 4 were never observed. Acetate production was measured for the cultures identified in region 2, and the acetate production is quantitatively compared to the in silico predictions in Fig. 6.

[0121] The optimal growth rate surface was constructed over the succinate-oxygen PhPP, and the measured flux vectors fell near the edge of the polytope that corresponded to the LO (Fig. 7). The flux vectors also identified a locus of points on the phase surface in region 2 with a constant oxygen uptake rate equal to the maximal oxygen uptake limit of the system. To quantitatively test the predictive capability of the in silico analysis and the in silico derived hypothesis, we employed a piecewise linear model to describe our hypothesis and the experimentally observed flux vectors. The piecewise linear model is defined as follows: we identified the locus of points defined by the flux vector for a range of succinate uptake rates and an oxygen uptake limit. Below the oxygen uptake limit, the locus of points lies along the LO, and above the oxygen uptake limit the locus of points lies along the phase surface with a constant oxygen uptake rate (the oxygen uptake limit). Based on the piecewise linear model, the succinate uptake rate was used to predict the oxygen uptake rate and the growth rate, and the other two permutations were

also considered. From this analysis an overall average error between the in silico predictions and the experimental data was 10.7%.

[0122] This Example shows that the strain used exhibited optimal aerobic growth using acetate and succinate as primary substrates. No adaptive evolution was necessary to achieve this optimal performance.

EXAMPLE 4

EVOLUTION OF A SUB-OPTIMAL *E. COLI* STRAIN TO OPTIMALITY

[0123] This Example demonstrates that *E. coli* can undergo some phenotypic adaptation from a sub-optimal growth state to an optimal state determined *in silico*.

[0124] **Glucose**: The glucose-oxygen PhPP contains six distinct regions (Figure 8) Like the succinate-oxygen PhPP, region 1 represents futile cycles and sub-optimal growth performance, whereas region 2 is characterized by acetate overflow metabolism. The two are separated by the LO.

[0125] As before, the cellular growth rate, OUR, and glucose uptake rate (GUR) were experimentally determined over a range of glucose concentrations and temperatures. Most experimentally determined values for the GUR, OUR, corresponded to points on the LO or slightly in region 2 of the PhPP (Figure 8), where the predicted acetate secretion was experimentally observed.

[0126] In three dimensions, the measured growth rates lie on the surface of the solution space near the edge corresponding to the LO, but are not tightly clustered there (Figure 9). We therefore kept the strain in sustained exponential growth (16) over a 40-day period (about 750 generations) using serial transfer under constant growth conditions to determine whether the metabolic phenotype would evolve (Figures 10 and 11). Fitness indeed increased, as shown by movement of the experimental points parallel to the LO, but there was no qualitative change in the phenotype.

EXAMPLE 5

EVOLUTION OF A SUB-OPTIMAL *E. COLI* STRAIN TO OPTIMALITY

[0127] This Example demonstrates that *E. coli* can undergo significant phenotypic adaptation from a sub-optimal growth state to an optimal state determined *in silico*.

[0128] **Glycerol**: The glycerol-oxygen PhPP consists of 5 regions with features resembling those seen in the PhPPs for acetate, succinate and glucose. In particular, a region with futile cycles (phase 1) is separated from an acetate overflow region (phase 2) by the LO.

[0129] The growth performance over a range of glycerol concentrations was experimentally determined as before. In sharp contrast to growth on malate or glucose, however, the experimental values for growth were scattered throughout phase 1, far from the LO (Figure 12) and the surface of optimality (Figure 13). Unlike the other substrates examined, glycerol thus supports only sub-optimal growth of *E. coli* K-12.

[0130] As before, we therefore performed a long-term adaptive growth experiment, this time using glycerol as the sole carbon source. The original strain was again kept in prolonged exponential growth for 40 days by serial transfer (17), maintaining a temperature of 30°C, a glycerol concentration of 2g/L, and sufficient oxygenation. Growth rate, glycerol uptake rate (GlUR) and OUR were determined every ten days.

[0131] A forty-day evolutionary path (E1) was traced in phase 1, eventually converging on the LO (Figure 14). During this period, the growth rate more than doubled from 0.23 hr^{-1} to 0.55 hr^{-1} (Figure 15). Further testing of the resulting evolved strain (which was frozen and stored) revealed higher specific growth rates and biomass yields than the parental strain. All of the data obtained fell on or near the LO as it had on the final day of the long-term culture (Figures 16 and 17), indicating that the evolved strain had attained an optimal growth performance on glycerol consistent with predictions *in silico*. A second, independent adaptation experiment gave a similar but not identical evolutionary trajectory (E2), converging near the same endpoint. *E. coli* can therefore undergo significant phenotypic adaptation from a sub-optimal growth state to an optimal state determined *in silico*.

[0132] Optimal growth performance of bacteria thus appears to conform with the predictions of *in silico* analysis. On some substrates, such as acetate and succinate, the cell may display optimal growth, whereas on others such as glucose and glycerol it may not. In the latter case growth evolves towards a phase plane predicted optimal performance.

[0133] All of the references cited herein are incorporated by reference. Although the invention has been described with reference to the above examples, it will be understood that modifications and variations are encompassed within the spirit and scope of the invention. Accordingly, the invention is limited only by the following claims.